Accumulations of Projections—A Unified Framework for Random Sketches in Kernel Ridge Regression: Supplementary Materials

Yifan Chen
UIUC
yifanc10@illinois.edu

Yun Yang
UIUC
yy84@illinois.edu

6 APPENDIX OUTLINE

This appendix is arranged as follows. In Section 7, we introduce some useful facts to help illustrate the assumptions in Theorem 1 in the main paper, and some matrix inequalities to prepare for the proof in Section 8. With those matrix inequalities, we prove Theorem 2 in the main paper, the condition to guarantee K-satisfiability with high probability. Finally, in Section 9, we provide more details on the experiments in the main paper, and some additional experiments to comprehensively compare our sketching method with the other candidates.

7 USEFUL FACTS

In this section, we will first introduce some preliminary knowledge of RKHS kernels, and then provide some useful matrix inequalities heavily used later.

7.1 Preliminaries

Theorem 1 in the main paper guarantees that for a K-satisfiable sketching matrix S, with high probability the approximation error $\|\hat{f}_S - \hat{f}_n\|_n^2$ would be bounded by $\lambda + \frac{d_\lambda}{n}$. The result of this theorem is powerful, while it relies on some assumptions on the kernel function and the sketching matrix used. The authors of the previous work (Liu et al., 2018) have summarized three assumptions for Theorem 1, and in this subsection, we provide the necessary introduction to the eigendecomposition of the kernel function $\mathcal{K}$, to ease the explanation of the assumptions.

In the preliminaries in the main paper, we have known that the associated kernel function $\mathcal{K}$ of an RKHS $\mathbb{H}$ is positive semi-definite. Further by Mercer's theorem, $\mathcal{K}$ has the following spectral expansion:

$$\mathcal{K}(x, x') = \sum_{i=1}^{\infty} \mu_i \phi_i(x)\phi_i(x'), \quad \forall x, x' \in \mathcal{X},$$

where $\mu_1 \geq \mu_2 \geq \cdots \geq 0$ are denoted as the eigenvalues of $\mathcal{K}$, and $\{\phi_i\}_{i=1}^{\infty}$ actually form a basis in $L^2(\mathcal{X})$, i.e.

$$\langle \phi_i, \phi_j \rangle_{L^2(\mathcal{X})} = \delta_{ij}, \quad \langle \phi_i, \phi_j \rangle_{\mathbb{H}} = \delta_{ij}/\mu_i.$$

The eigenvalues $\{\mu_i\}_{i=1}^{\infty}$ of the kernel function $\mathcal{K}$ are closely related to the eigenvalues $\{\sigma_i\}_{i=1}^{n}$ of the rescaled empirical kernel matrix $\frac{1}{n}K$. The eigenvalue pair (μ_i, σ_i) of the same index i would roughly have the same magnitude, and more details can be found in the work (Braun, 2006).

7.2 Useful Matrix Inequalities

The key step in the proof of K-satisfiability is to control the operator norm of a random matrix. Here we give a sequence of matrix inequalities used later in Section 8. More details and proofs of those matrix inequalities could be found in the note (Tropp, 2012).

Theorem 3 (Matrix Chernoff)**.** *Consider a finite sequence $\{X_k\}$ of independent, random, self-adjoint matrices with dimension n. Assume that each random matrix satisfies*

$$X_k \succcurlyeq \mathbf{0} \quad \text{and} \quad \lambda_{\max}(X_k) \leq R \quad \text{almost surely.}$$

Define

$$\mu_{\max} := \lambda_{\max}\left(\sum\nolimits_k \mathbb{E}\, X_k\right).$$

Then for $\delta > 0$,

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k X_k\right) \geq (1+\delta)\mu_{\max}\right\} \leq n \cdot \left[\frac{\mathrm{e}^{-\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_{\max}/R}.$$

The matrix Chernoff inequalities describe the behavior of a sum of independent, random, positive-semidefinite matrices. Specifically, they are well suited to study the operator norm of an arbitrary random matrix A with independent columns a_k's, due to the fact $\|A\|^2 = \|AA^T\| = \|\sum_k a_k a_k^T\|$, and the property that $(a_k a_k^T)$'s are independent, random, self-adjoint matrices.

We also introduce Bernstein matrix inequality to show the normal concentration near the mean of the random matrices.

Theorem 4 (Matrix Bernstein)**.** *Consider a finite sequence $\{X_k\}$ of independent, random, self-adjoint matrices with dimension n. Assume that each random matrix satisfies*

$$\mathbb{E}\, X_k = \mathbf{0} \quad \text{and} \quad \|X_k\| \leq R \quad \text{almost surely.}$$

Then, for all $t \geq 0$,

$$\mathbb{P}\left\{\|\sum\nolimits_k X_k\| \geq t\right\} \leq 2n \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right) \quad \text{where} \quad \sigma^2 \geq \left\|\sum\nolimits_k \mathbb{E}\left(X_k^2\right)\right\|.$$

Similar to the ordinary Bernstein inequality for random variables, the decaying rate of the tail of the sum would be determined by the variance of the matrix sum and the uniform bound on the maximum eigenvalue of each summand. As in our framework, each column of S is also an accumulation of m independent columns, we further introduce a rectangular version of the matrix Bernstein inequality, which is an immediate corollary of Theorem 4.

Theorem 5 (Matrix Bernstein: Rectangular Case)**.** *Consider a finite sequence $\{Z_k\}$ of independent, random matrices with dimensions $n_1 \times n_2$. Assume that each random matrix satisfies*

$$\mathbb{E}\, Z_k = \mathbf{0} \quad \text{and} \quad \|Z_k\| \leq R \quad \text{almost surely.}$$

Define

$$\sigma^2 := \max\left\{\left\|\sum\nolimits_k \mathbb{E}(Z_k Z_k^*)\right\|,\ \left\|\sum\nolimits_k \mathbb{E}(Z_k^* Z_k)\right\|\right\}.$$

Then, for all $t \geq 0$,

$$\mathbb{P}\left\{\left\|\sum\nolimits_k Z_k\right\| \geq t\right\} \leq (n_1 + n_2) \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right).$$

8 MISSING PROOFS

In this section, we will present the proof for Theorem 2 in the main paper.

Proof. The main idea of the proof is to utilize Theorem 5, the rectangular version matrix Bernstein inequality, to control the upper bound of the column ℓ_2 norm in the sketching matrix S, and then further address the properties of the whole matrix S.

We start with the first condition in K-satisfiability. Observing the form of the matrix $U_1^T SS^T U_1 - I$, we find its expectation is zero and thus matrix Bernstein inequality (Theorem 4) can be applied. Before that, we need to give a high probability operator norm upper bound R for the summand $U_1^T S_k S_k^T U_1 - I/d$, which can be derived from the norm upper bound for the vector $U_1^T S_k$.

Based on Algorithm 1 in the main paper, S_k can be decomposed as

$$S_k = \sum_{j=1}^{m} \frac{1}{\sqrt{m}} S_{k,j}, \quad \forall k \in [d],$$

where $S_{k,j}$ is a single sub-sampling column which would be $\frac{1}{\sqrt{dp_i}} e_i$ with probability p_i. We take $\frac{1}{\sqrt{m}} U_1^T S_{k,j}$'s as the random matrices and apply Theorem 5 to them. We then need to specify the parameters R and σ^2 in Theorem 5. Here we let $R = \sqrt{\frac{2M}{md}}$, as $M \geq \max_i \frac{\|\tilde{\psi}_i\|^2}{p_i} \geq \max_i \frac{\|\tilde{u}_i\|^2}{2p_i}$, where $\tilde{u}_i$ is the sub-vector of the first d_δ elements in the i-th column u_i of U. We can verify $R \geq \|\frac{1}{\sqrt{m}} U_1^T S_{k,j}\|, \forall j \in [m]$. Another parameter $\sigma^2 = \frac{d_\delta}{d}$ is induced by the fact $\mathbb{E}\, S_{k,j} S_{k,j}^T = \frac{1}{d} I$.

By Theorem 5 we have the following probability inequality

$$\mathbb{P}\left\{\|U_1^T S_k\| > t\right\} \leq (d_\delta + 1) \exp(\frac{-t^2/2}{\frac{d_\delta}{d} + \sqrt{\frac{2M}{md}} t/3}),$$

and further by union bound we obtain

$$\mathbb{P}\left\{\max_{k \in [d]} \|U_1^T S_k\| > t\right\} \leq d(d_\delta + 1) \exp(\frac{-t^2/2}{\frac{d_\delta}{d} + \sqrt{\frac{2M}{md}} t/3})$$

$$\mathbb{P}\left\{\max_{k \in [d]} \|U_1^T S_k\| > t = \frac{1}{3} uR + \sqrt{\frac{1}{9} u^2 R^2 + 2u\sigma^2}\right\} \leq d(d_\delta + 1) \exp(-u),$$

where we substitute u for t in the last inequality to better control the probability. After the substitution, we upper bound the right hand side by $\rho/4$, and have $u \sim \log \frac{n}{\rho}$. Thus with probability $1 - \frac{\rho}{4}$, we upper bound the vector norm $\|U_1^T S_k\|$ by t, and further bound the norm of the zero mean matrix $\|U_1^T S_k S_k^T U_1 - \frac{1}{d} I\|$ by t^2.

For the simplicity of notation, we would denote $U_1^T S_k S_k^T U_1 - \frac{1}{d} I$ as X_k in this paragraph. We still need to control $\|\mathbb{E}\, X_k^2\|$ to apply Theorem 4. We first expand $\mathbb{E}\, X_k^2$ as

$$\mathbb{E}\, X_k^2 = \mathbb{E} \sum_{i_1, i_2, i_3, i_4 \in [n]} (\frac{Z_{i_1} Z_{i_2}}{md} - \frac{\delta_{i_1 i_2}}{d})(\frac{Z_{i_3} Z_{i_4}}{md} - \frac{\delta_{i_3 i_4}}{d}) \tilde{u}_{i_1} \tilde{u}_{i_2}^T \tilde{u}_{i_3} \tilde{u}_{i_4}^T, \tag{4}$$

where $S_k^T = \frac{1}{\sqrt{md}}(Z_1, Z_2, \cdots, Z_n)$, and further $Z_i = \frac{1}{\sqrt{p_i}} \sum_{j=1}^m Z_{ij}$, where Z_{ij}'s are i.i.d and they would be ± 1 with probability $\frac{p_i}{2}$ respectively, or be 0 with probability $1 - p_i$. The setting represents the fact that S_k is the accumulation of m independent columns $\frac{1}{\sqrt{m}} S_{k,j}$. By some calculation, we have $\mathbb{E}\, Z_{i_1} Z_{i_2} = m\delta_{i_1 i_2}$, and for most combination of i_1, i_2, i_3, i_4 the summand in equation (4) would be zero. To exactly compute equation (4), we only need to consider the following four cases:

1. $i_1 = i_2 = i_3 = i_4$,

2. $i_1 = i_2 \neq i_3 = i_4$,

3. $i_1 = i_3 \neq i_2 = i_4$,

4. $i_1 = i_4 \neq i_2 = i_3$.

For the first case, we have

$$\mathbb{E}\sum_{i_1=i_2=i_3=i_4}(\frac{Z_{i_1}^2}{md}-\frac{1}{d})^2\tilde{u}_{i_1}\tilde{u}_{i_1}^T\tilde{u}_{i_1}\tilde{u}_{i_1}^T=\frac{1}{d^2}\sum_{i=1}^{n}(\frac{1}{mp_i}+\frac{3(m-1)}{m}-1)\tilde{u}_{i_1}\tilde{u}_{i_1}^T\tilde{u}_{i_1}\tilde{u}_{i_1}^T;$$

for the second case we have

$$\mathbb{E}\sum_{i_1=i_2\neq i_3=i_4}(\frac{Z_{i_1}^2}{md}-\frac{1}{d})(\frac{Z_{i_3}^2}{md}-\frac{1}{d})\tilde{u}_{i_1}\tilde{u}_{i_1}^T\tilde{u}_{i_3}\tilde{u}_{i_3}^T=-\frac{1}{md^2}\sum_{i_1\neq i_3}\tilde{u}_{i_1}\tilde{u}_{i_1}^T\tilde{u}_{i_3}\tilde{u}_{i_3}^T;$$

for the third case we have

$$\mathbb{E}\sum_{i_1=i_3\neq i_2=i_4}(\frac{Z_{i_1}Z_{i_2}}{md})^2\tilde{u}_{i_1}\tilde{u}_{i_2}^T\tilde{u}_{i_1}\tilde{u}_{i_2}^T=\frac{m-1}{md^2}\sum_{i_1\neq i_2}\tilde{u}_{i_1}\tilde{u}_{i_2}^T\tilde{u}_{i_1}\tilde{u}_{i_2}^T;$$

for the fourth case we have

$$\mathbb{E}\sum_{i_1=i_4\neq i_2=i_3}(\frac{Z_{i_1}Z_{i_2}}{md})^2\tilde{u}_{i_1}\tilde{u}_{i_2}^T\tilde{u}_{i_2}\tilde{u}_{i_1}^T=\frac{m-1}{md^2}\sum_{i_1\neq i_2}\tilde{u}_{i_1}\tilde{u}_{i_2}^T\tilde{u}_{i_2}\tilde{u}_{i_1}^T.$$

Combining the pieces together, we obtain

$$\mathbb{E}\,X_k^2=\frac{1}{md^2}\sum_{i=1}^{n}\frac{\|\tilde{u}_i\|^2}{p_i}\tilde{u}_{i_1}\tilde{u}_{i_1}^T+\frac{m-1}{md^2}d_\delta I+\frac{m-2}{md^2}I,$$

which implies

$$\|\,\mathbb{E}\,X_k^2\|\leq\frac{1}{d^2}(\frac{2}{m}M+\frac{m-1}{m}d_\delta+\frac{m-2}{m}),$$

and hence we let $\|\sum_k\mathbb{E}\,X_k^2\|\leq\sigma_b^2:=\frac{1}{d}(\frac{2}{m}M+d_\delta+1)$.

Applying Theorem 4, we have the following probability inequality:

$$\mathbb{P}\left\{\|U_1^TSS^TU_1-I_{d_\delta}\|_{op}\geq 1/2\right\}\leq 2d_\delta\exp(\frac{-1/8}{\sigma_b^2+\frac{1}{6}t^2})\leq\frac{\rho}{4}.$$

To make the last inequality hold, we need

$$\sigma_b^2+\frac{1}{6}t^2\lesssim 1/\log\frac{n}{\rho},$$

which turns out to imply

$$d\gtrsim d_\delta\log^2(\frac{n}{\rho})$$
$$md\gtrsim M\log^3(\frac{n}{\rho}).$$

To complete the proof for the theorem, we still need to validate the requirements on m, d above can induce the second condition in K-satisfiability. We first rewrite the target $\|S^TU_2\Sigma_2^{\frac{1}{2}}\|$ as $\|\overline{(\Sigma+\delta I)}^{\frac{1}{2}}\Psi_\delta S\|$, where $\overline{(\Sigma+\delta I)}^{\frac{1}{2}}$ means the matrix $\Sigma+\delta I$ would set the first d_δ diagonal elements as zero and only keep the rest ones.

As above we start with the control over the maximal norm of the column $\overline{(\Sigma+\delta I)}^{\frac{1}{2}}\Psi_\delta S_k$. For simplicity we will reuse the notation σ^2, R to apply Theorem 5. Recall $S_k=\sum_{j=1}^m\frac{1}{\sqrt{m}}S_{k,j}$, this time $\|\frac{1}{\sqrt{m}}\overline{(\Sigma+\delta I)}^{\frac{1}{2}}\Psi_\delta S_{k,j}\|$ is bounded by

$$\frac{1}{\sqrt{md}}\sqrt{2\delta}\sqrt{\max_i\frac{\|\psi_i\|^2-\|\tilde{\psi}_i\|^2}{p_i}}\leq R:=\sqrt{\frac{2\delta M}{md}},$$

and

$$\mathbb{E}\sum_j \frac{1}{m} S_{k,j}^T \Psi_\delta^T \overline{(\Sigma+\delta I)} \Psi_\delta S_{k,j} = \frac{\mathrm{Tr}(\Sigma_2)}{d} \leq \sigma^2 := \frac{C d_\delta}{d}\delta,$$

in which the last inequality is induced by the fast eigenvalue decay rate in Assumption 1 (Liu et al., 2018, Lemma 3.1). Again by Theorem 5 and the union bound, we have

$$\mathbb{P}\left\{\max_{k\in[d]} \|\Sigma_2 U_2^T S_k\| > t = \frac{1}{3}uR + \sqrt{\frac{1}{9}u^2R^2 + 2u\sigma^2}\right\} \leq d(n+1)\exp(-u) \leq \rho/4.$$

Given the high probability norm upper bound t, to apply Theorem 3 we re-define $X_k = \overline{\Sigma}^{\frac{1}{2}} U^T S_k S_k^T U \overline{\Sigma}^{\frac{1}{2}}$, and $\mu_{\max} = \lambda_{\max}\left(\sum_k \mathbb{E}\, X_k\right) = \delta$. We finally have

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k X_k\right) \geq (1+1)\mu_{\max}\right\} \leq n \cdot \left[\frac{\mathrm{e}^{-1}}{4}\right]^{\mu_{\max}/t^2} \leq \rho/4.$$

To make the second condition in K-satisfiability hold we only need to validate

$$t^2/\delta \lesssim 1/\log\frac{n}{\rho},$$

which is actually satisfied by the derived requirements on m, d above. ♢

9 MORE ON SIMULATIONS

In this section, we provide the complete experiment settings and several additional figures to further illustrate our method.

9.1 Experiment Settings in Figure 1 in the Main Paper

This figure is only shown for illustration, and the settings are relatively simple. In Figure 1 we consider three representative sketching methods, the Gaussian sketching method, the classical Nyström method, and our accumulation method with $m = 5$. We compare the estimation error $\|\hat{f}_S - \hat{f}_n\|_n^2$ of each other and also report the total runtime in Figure 1.

Specifically, we run the experiment on a bimodal distribution over $\mathbb{R}^3$. The bimodal distribution has two components: with probability $\frac{n}{n+n^\gamma}$ generating a Unif$[0,1]^3$; and with probability $\frac{n^\gamma}{n+n^\gamma}$ generating a random variable with pdf $\prod_{j=1}^3(5-2x_j)$ for $x_j \in [2, 2.5]$, where n is the sample size varying from $n = 1,000$ to $16,000$ and $\gamma = 0.5$. In addition, by cross validation the Matérn kernel with smoothness parameter $\nu = 0.5$ is chosen, and the regularization parameter of the KRR λ is set as $0.3n^{-\frac{4}{7}}$. The true regression function we use is $f^*(x) = g(\|x\|/3)$, with

$$g(x) = 1.6|(x-0.4)(x-0.6)| - x(x-1)(x-2) - 0.5,$$

and i.i.d. noises follow $\mathcal{N}(0, 0.25)$. We use uniform sub-sampling distribution for all the applicable methods, and the projection dimension d is chosen as $\lfloor 1.3n^{\frac{3}{7}} \rfloor$. The results finally reported in Figure 1 are averaged over 30 replicates.

9.2 Experiment Settings in Figure 2 in the Main Paper

In this experiment, we compare the approximation error $\|\hat{f}_S - \hat{f}_n\|_n^2$ among the KRR estimators obtained by the sketching matrices with different m, including the Gaussian sketch as an instance of $m = \infty$. As above, we run the experiment on a bimodal distribution over $\mathbb{R}^3$. The bimodal distribution has two components: with probability $\frac{n}{n+n^\gamma}$ generating a Unif$[0,1]^3$; and with probability $\frac{n^\gamma}{n+n^\gamma}$ generating a random variable with pdf $\prod_{j=1}^3(5-2x_j)$ for $x_j \in [2, 2.5]$, where n is the sample size varying from $n = 1,000$ to $8,000$ and $\gamma = 0.6$. In addition, by cross validation the Gaussian kernel with bandwidth $\sigma = 1.5n^{-\frac{1}{7}}$ is chosen, and the regularization parameter of the KRR λ is set as $0.5n^{-\frac{4}{7}}$. The true regression function f^* we use is $f^*(x) = g(\|x\|/3)$, with

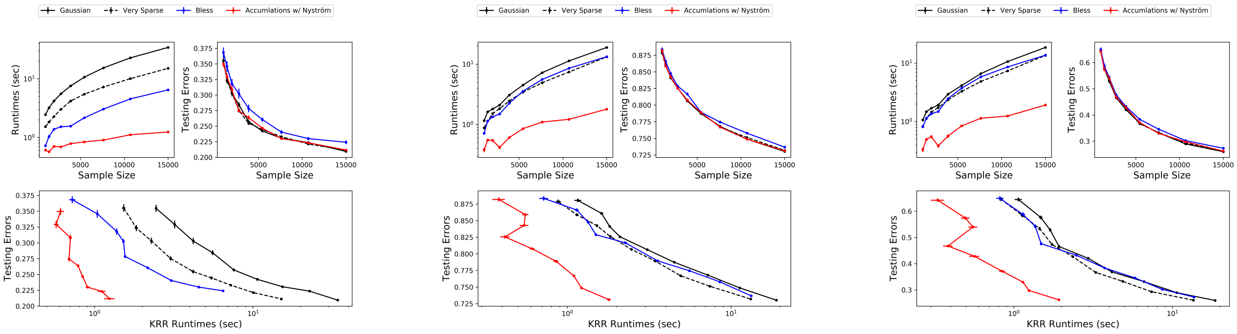
$$g(x) = 1.6|(x - 0.4)(x - 0.6)| - x(x - 1)(x - 2) - 0.5,$$

and i.i.d. noises follow $\mathcal{N}(0, 0.25)$. We use uniform sub-sampling distribution for all the applicable methods, and the projection dimension d varies from $\lfloor 0.3n^{\frac{3}{7}} \rfloor$ to $\lfloor 3n^{\frac{3}{7}} \rfloor$. The results reported in Figure 2 are averaged over 30 replicates.

9.3 Additional Experiments and Experiment Settings in Figure 3 in the Main Paper

As mentioned in the main paper, we evaluate the comprehensive performance of our accumulation method on three real datasets and also consider the usage of Falkon (Rudi et al., 2017). Due to the space limit in the main paper, we move the complete experiment settings and results to this section and will demonstrate them as follows, including the settings in Figure 3 in the main paper.

We verify the effect of our method on three datasets, `RadiusQueriesAggregation` (Savva et al., 2018; Anagnostopoulos et al., 2018)(denoted by RQA), `CASP` (Dua and Graff, 2017), and `PPGasEmission` (KAYA et al., 2019)(denoted by GAS), all downloaded from the UCI ML Repository (Dua and Graff, 2017). For those datasets, RQA contains 200,000 data points and 4 features; CASP contains 45,730 data points and 9 features; GAS contains 36,733 data points and 10 features; in this subsection we use d_X to represent the number of the features. To show the evolving trend, we set a sequence of sample sizes beforehand (from 1,000 to 15,000, at the limit of our computational feasibility) and in each round run the experiment on a subset of the whole data points with the given sample size n. The testing errors are estimated on a random subset (20% of the original dataset) which is not used in the training. We begin by normalizing the features to have variance 1 in the randomly drawn dataset, before obtaining the empirical kernel matrix using Matérn kernel (the smoothness parameter $\nu = 1.5$). The regularization parameter λ of KRR is $0.9 \cdot n^{-\frac{3+d_X}{3+2d_X}}$. We set the projection dimension as $\lfloor 1.5 \cdot n^{\frac{d_X}{3+2d_X}} \rfloor$ for all sketching methods. The candidate methods include the Gaussian sketching method, very sparse random projection (Li et al., 2006), the Nyström method with Bless (Rudi et al., 2018), and our accumulation method with $m = 4$. Among all the experiments, the number of sub-samples used in Bless is chosen as $\lfloor 3 \cdot n^{\frac{d_X}{3+2d_X}} \rfloor$. The usage of Falkon (Rudi et al., 2017) is also considered, and we provide some experimental results to show our method still attain the optimal trade-off between statistical accuracy and computational efficiency. All the results reported below in Figure 4 (in which the first subplot is Figure 3 in the main paper) and Figure 5 are averaged over 30 replicates.



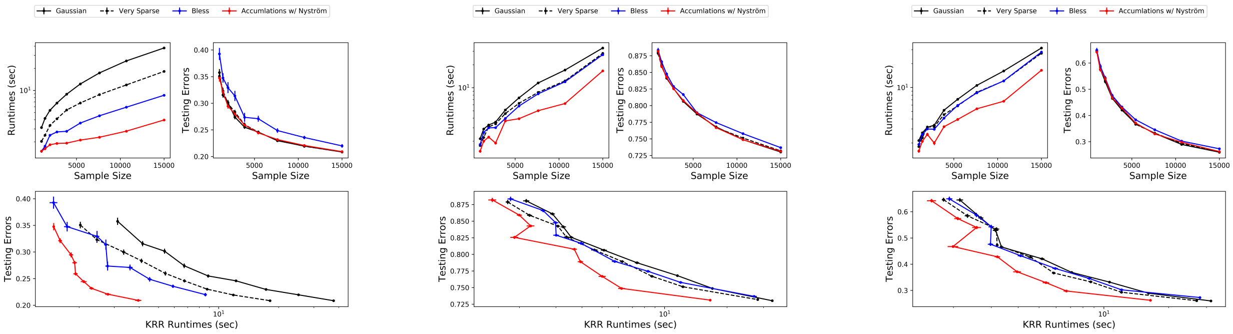
(a) RQA

(b) CASP

(c) GAS

Figure 4: Trade-off between Accuracy and Efficiency without Falkon

Basically, those experiments demonstrate that in practice, a medium m could substantially improve the accuracy of the classical Nyström method with uniform sub-sampling distribution, and the extra cost is much lower compared to other advanced methods. Specifically, all the methods above can be combined with a fast KRR solving method Falkon. Under our experiment settings, Falkon maintains the estimation accuracy, while not accelerate the training much since the sample size is not large enough. As shown in Figure 5, the usage of Falkon



(a) RQA (b) CASP (c) GAS

Figure 5: Trade-off between Accuracy and Efficiency with Falkon

would not change the main conclusion that our accumulation method (the red solid curve) provides the optimal trade-off between statistical accuracy and computational efficiency, among all the candidate sketching methods.

References

Anagnostopoulos, C., Savva, F., and Triantafillou, P. (2018). Scalable aggregation predictive analytics. *Applied Intelligence*, 48(9):2546–2567.

Braun, M. L. (2006). Accurate error bounds for the eigenvalues of the kernel matrix. *Journal of Machine Learning Research*, 7(Nov):2303–2328.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

KAYA, H., TÜFEKCİ, P., and UZUN, E. (2019). Predicting CO and NOxemissions from gas turbines: novel data and abenchmark PEMS. *TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES*, 27(6):4783–4796.

Li, P., Hastie, T. J., and Church, K. W. (2006). Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–296.

Liu, M., Shang, Z., and Cheng, G. (2018). Nonparametric testing under random projection. *arXiv preprint arXiv:1802.06308*.

Rudi, A., Calandriello, D., Carratino, L., and Rosasco, L. (2018). On fast leverage score sampling and optimal learning. In *Advances in Neural Information Processing Systems*, pages 5672–5682.

Rudi, A., Carratino, L., and Rosasco, L. (2017). Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pages 3888–3898.

Savva, F., Anagnostopoulos, C., and Triantafillou, P. (2018). Explaining aggregates for exploratory analytics. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 478–487. IEEE.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434.